DOCUMENT RESUME

ED 099 429 95 TM 004 307

AUTHOR Ebel, Robert L.
TITLE State Testing Programs: Status, Problems, and Prospects. TM Report 40.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ETS-TM-40
PUB DATE Dec 74
CONTRACT OEC-0-70-3797-519
NOTE 6p.; For related documents, see ED 086 721 and 087 789

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Criterion Referenced Tests; *Educational Assessment; *Educational Testing; Standardized Tests; *State Programs; State Surveys; *Surveys; *Testing Programs
IDENTIFIERS Tailor Made Tests

ABSTRACT

The current status of state testing programs is assessed drawing primarily on information provided by the Educational Testing Service publication, "State Testing Programs, 1973 Revision." Increases in state operated programs are indicated and are probably due to an increase in federal money for testing purposes. Because of possible confusion over the differences between a state testing program, a state assessment program, and a state testing service, some explanation is given as to the properties of each. A history of state testing programs is outlined, and new directions for such programs are proposed. Criterion-referenced and norm-referenced testing is contrasted, and the advantages and limitations of criterion-referenced tests are indicated. The problem of evaluating affective educational outcomes is explored and may be explained by the very limited role of noncognitive tests in state testing programs. The relation between the purposes of testing and the time of year the tests are given is discussed, and this timing is seen to affect the extent to which a particular purpose is served well or poorly. As to the type of test that should be given, standardized tests and tailor-made tests are compared, and their advantages and limitations are discussed. (RC)

ED 092 029

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 40 DECEMBER 1974

STATE TESTING PROGRAMS: STATUS, PROBLEMS, AND PROSPECTS

Robert L. Ebel

The Current Status of State Testing Programs

State programs of testing and assessment are prominent features of the contemporary educational scene. In a few cases, these programs simply continue efforts to measure pupil achievements, efforts that began decades or, in one case, over a century ago. In many more cases, they are recent innovations, responses to increasing demands for accountability, or to needs for the evaluation of innovations in education.

Only 13 of the 50 states do not now have a statewide testing program. In three of these, plans are being developed for the inauguration of a testing program. Seven states in the midwest have programs operated by agencies of their state universities. These are supported mainly out of local school district budgets. While participation in the programs is voluntary, many, or in some cases most, of the schools in the state take advantage of the testing services offered by the universities.

In 31 states[1], testing programs are operated under the direction of the state department of education. Nineteen of these states report that their testing programs are substantially or totally supported by the federal government from funds available under Title III of the Elementary and Secondary Education Act. In the other 14 states, the necessary funds are provided by the state government.

A Survey of State Testing Programs

These and a number of other interesting and useful facts are presented in *State Testing Programs, 1973 Revision*, a publication developed by Educational Testing Service (ETS) in collaboration with the Conference of Directors of State Testing Programs. The factual material presented in that report was obtained through telephone interviews with the individual or individuals in each state who appeared most likely to provide detailed and accurate information. These state authorities were told in advance what questions would be asked during the interview so that they could be prepared to give accurate answers.

[1]In one state, both the state department of education and a state university operate testing programs.

Then and Now in State Testing Programs

The immediate forerunner of the 1973 publication was a similar report prepared by ETS in 1968. A much earlier publication, intended to serve the same purpose, was *State Testing and Evaluation Programs* by David Segel, published in 1951 by the U.S. Office of Education. It is interesting to see how many states were offering each kind of program then and now.

Program operated by	*1951*	*1973*
State department	17	31
State university	19	7
No program	17	13
	53[2]	51[3]

Note the sharp increase in the number of testing programs operated by state departments of education. Part of this increase, perhaps most of it, is almost certainly due to the availability of federal funds (e.g. NDEA-1958, ESEA-1965) for testing purposes. When those funds are no longer available, the number of state operated programs may be cut drastically.

Note also the sharp decrease in the number of testing programs operated by state universities. These programs are usually voluntary since the universities have no direct control over local education authorities. Their costs are born by the local district. They tend to emphasize guidance and instructional assistance rather than assessment of educational effectiveness. As state legislatures make increasingly heavy investments in local school support, they become increasingly interested in the kind and amount of education their dollars are buying. These may be some of the factors which account for the replacement of voluntary, university-based testing pro-

[2]In 1951, five states had dual testing programs, one operated by the state department and the other by the state university.

[3]In 1973 only one state had both state department- and university-operated programs.

TM 004 607

This publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, represent official National Institute of Education position or policy.

grams by mandatory programs operated out of state departments of education.

The figures just presented and discussed probably convey a reasonably accurate picture of the situation then and now with respect to state testing programs. However, despite the care taken by ETS to obtain accurate information, there may have been some differences of opinion among respondents to the survey over the essential characteristics of a "state testing program." Does the term apply only to programs operated under mandate of the state government? Does it apply only to programs in which participation of all local education authorities is required? Does it apply only to programs in which all examinees are required to take the same tests? Does it apply to tests given to evaluate the effectiveness of educational programs? Is assessment the same as testing? Questions such as these suggest that it may be useful to define, and to distinguish among the terms:

1. State testing program
2. State assessment program
3. State testing service

Testing, Assessment, and Service Programs

A state testing program is available state-wide but is limited to the schools in a particular state. It involves the use of a common test or set of tests. Participation in the program may be mandatory or voluntary. Program costs may be borne by a state agency or by the local school. The organization responsible for administration of the testing program provides standard directions for scheduling and administering tests. It also provides assistance in obtaining, interpreting, and utilizing the test scores.

A state assessment program is similar to a state testing program and may, in some cases, be identical. That is, some testing programs may be called assessment programs mainly to avoid the threatening or otherwise unpleasant connotations of the term "testing." But there may be some characteristic differences.

An assessment program is likely to focus more on the effectiveness of an educational program than on the achievements of individual pupils. Participation in an assessment program is more likely to be mandatory, and the costs are more likely to be borne by a state department.

The testing itself may involve matrix sampling. That is, both pupils and items are sampled. Instead of asking all pupils to take the same long, comprehensive test, different pupils take different, much shorter, sets of test items. This means a good estimate of group achievement can be obtained in much less time of testing.

Despite these differences, it seems reasonable to regard a state assessment program as one kind of state testing program. Surely it is useful to include data on such programs in a survey of the kind recently made by ETS, and it was clearly the intention of ETS to include them. But it is equally clear that at least one state with an extensive program of assessment did not report it as a testing program.

A state testing service is usually operated to assist local schools in (1) obtaining the tests they want to use, (2) scoring the tests, and (3) interpreting and utilizing the test results. It operates statewide but involves no limited or prescribed set of tests. Most of the costs of testing are borne by the local school districts. In some states, an annual conference is used as a kind of inservice toward more effective test utilization.

Clearly, there are important differences between a state testing program and a state testing service. But again, it seems useful to include reports on such services in any general survey of state testing programs.

A Little History

State testing programs have a long history. One, instituted by the Regents of the University of the State of New York, began in 1865. Later, some states began to administer tests to certify satisfactory completion of the first eight grades. But the rapid expansion of state testing programs, begun after World War I, was influenced by at least two factors. One was a general concern for efficiency in business, industry, and indeed all enterprises, including education. To determine efficiency one must measure results. The other factor was development and refinement of techniques for measuring educational achievements and psychological traits. This influence began well before World War I, but it was given strong impetus by needs for testing in the personnel selection and training programs of the military services.

The testing programs initiated in the 1920s and 1930s flourished for several decades. A few have continued to flourish. Others declined and were abandoned. One reason for this may be the basic antipathy of school administrators to external evaluations. Another may be that local schools, having developed their own special testing programs, see less need for participation in a uniform, external, state program. Still another may be the view strongly held by some educators that schools should be more concerned with a pupil's feelings, self-concept, and adjustment than with his knowledge, self-discipline, and achievement. Those who endorse this view regard tests and testing programs not as useful educational tools but as obstacles to attainment of the goals they seek. Finally, it is possible that some state testing programs languished and died simply because they were not good enough, because they did not seem to meet basic educational needs, as those needs were perceived by persons who controlled the schools.

Whatever the cause, the years since 1950 have witnessed a decline in the kind of state testing programs that flourished before 1950. But in the decade of the

sixties, a new set of forces began to operate to create a new set of state testing programs.

New Directions

One of these forces, probably the most powerful one, has already been mentioned. It is the increasing role of state legislatures in providing funds for local school operations. Their responsibility to the electorate is to see that these funds are well spent. Hence, they support testing programs that promise to provide some of the evidence of educational outcomes they want.

A second force is increasing skepticism concerning the effectiveness of contemporary schools. The alleged failure of inner city schools has been well advertised. Innovative programs like Head Start, designed and promoted by educators, have yielded disappointing results. There is a widespread feeling that schools could do a better job if they would only try harder. Performance contracting and other strategies to make the schools more accountable have had considerable appeal. The crucial role of testing in these strategies has lent support to the extension of state testing programs.

A third force contributing to the renaissance of state testing programs is the "new look" of testing. This goes beyond substitution of the word "assessment" for "testing" in the program designation. It goes beyond a shift in the focus of attention from individual pupil achievement to curricular and instructional effectiveness. It involves mainly a somewhat different approach to the measurement of achievement, an approach that has been designated by terms such as "content-referenced testing," "domain-referenced testing" or "criterion-referenced testing." Two of the programs reported in the recent ETS survey mention their use, or intended use, of criterion-referenced tests. Others are no doubt also using them or considering their use.

Criterion-Referenced Testing

Criterion-referenced testing is often contrasted with norm-referenced testing. The aim of the first mentioned is to determine how many, and which ones, of a specified set of instructional objectives have been attained. Thus, the result of such a test may be a number, a percent, or a list of attainments. The aim of the norm-referenced test, on the other hand, is to indicate how the attainments of a particular pupil compare with those of his peers. The results (raw scores) of norm-referenced tests are usually converted into percentiles, grade equivalents, or standard scores. The meanings of all of these converted scores are essentially relative.

Criterion-referenced tests have some obvious advantages over norm-referenced tests. They can indicate directly what, and how much, the learner knows and can do. In tightly structured sequential learning, they can indicate when the student is ready to move ahead to the next phase. And they help to avoid direct comparisons of one pupil's achievements with those of another. Such comparisons, often made unfairly, have been the basis for some criticisms of norm-referenced tests.

But there are also some possible limitations of criterion-referenced tests. Because they sometimes focus on the attainment of a limited number of separate, discrete, highly specific objectives, they may induce teachers to neglect cultivation of more general capabilities for dealing with other related but unspecified problems. They may not encompass adequately the very large number of interrelated concepts, facts, ideas, principles, meanings, understandings, and so on that constitute learning in many areas. Emphasis on discrete specifics may lead to neglect of the integration of ideas that gives unity and solidarity to a subject. It may cause teachers and students to seek adequate performance of specified tasks through sheer memorization or habit forming, at the expense of understanding.

Another possible limitation lies in the difficulty of specifying adequacy of performance (mastery?) with respect to each objective. Learning is almost always a matter of degree. The statement "You either know something or you don't!" does not describe accurately the acquisition of knowledge in most areas of learning. Nor would a similar statement describe accurately the acquisition of an ability. This leaves the assessor with the question "How much knowledge or ability is enough?" It is a question that can seldom be answered on other than an arbitrary, conventional, not-clearly-rational-or-defensible basis.

Related to this is the difficulty of determining reliably whether a particular student has achieved a particular objective. In many criterion-referenced tests the attainment of each objective is tested by only a few items, sometimes only by one. Single test items, or very short tests, are notoriously unreliable. As a consequence of this unreliability, a substantial number of the students tested may be judged wrongly to have attained, or to have failed to attain, an adequate level of achievement with respect to a particular objective.

With more widespread usage, and more varied experience, an answer may gradually emerge to the question "Do the practical advantages of criterion-referenced tests outweigh their practical limitations?" Thus far they seem to have been used most successfully in testing for acquisition of basic skills in the early elementary grades. Whether they can be used effectively at higher levels of education also remains to be seen.

It may be worth mentioning that relatively few well informed, penetrating analyses of the strengths and weaknesses of criterion-referenced and of norm-referenced tests have been published in educational journals. Specialists in educational measurement seem much more concerned with adapting the statistics of

norm-referenced tests to the somewhat different materials and purposes of criterion-referenced tests. This may be understandable, but it may also be regrettable.

Testing Noncognitive Educational Outcomes

It is generally agreed that while schools seldom succeed brilliantly in achieving their cognitive goals, they have distinctly better success in the cognitive than in the affective domain. It is apparently much more difficult to define objectives, develop instructional programs, and evaluate affective outcomes than cognitive ones.

These difficulties may be largely responsible for the very limited role of noncognitive tests in state testing programs. In only nine states are noncognitive areas tested. Two states give noncognitive tests at both elementary and secondary levels; four give them only to elementary school students, and three give them only to secondary school students.

Noncognitive areas most frequently tested in elementary schools are attitudes toward school and self-concept. Those most frequently tested in secondary schools are interests and attitudes toward school. There is a notable absence of any attempt to assess student values (apart from interests and attitudes) in any of the state testing programs.

Are cognitive outcomes being overemphasized in school programs and in the testing programs designed to measure their effectiveness? Some critics of contemporary education contend that they are. It seems to them that a person's interests and values, his aspirations and attitudes, and his self-concept are crucially important in determining the quality of life he will live, and his success in living it. They conclude that schools should stress the attainment of noncognitive goals fully as much as cognitive goals are now being stressed.

Other educators are inclined to question that conclusion. They do not minimize the importance of interests and aspirations, of attitudes and values. They do not object to the use of school time to consider cognitive aspects of these affective manifestations. But they contend that the school is not authorized, or equipped, to mold a student's values, his attitudes, his interests, or his aspirations to fit some prescribed specifications. They believe that it is largely inappropriate for schools to define goals, design treatments, and assess outcomes in these areas.

Legislation is pending in at least one state (Michigan) that expressly forbids schools or teachers from attempting to "educate" their students affectively, "... by acting as a change agent of attitudes, values, and religious or political beliefs of the pupils."[4] There is no doubt that students *do* acquire affective responses in school, as they do at home and elsewhere. The question is whether or not schools should set out to teach certain affective responses purposefully and directly and whether they have any socially acceptable, noncognitive means for doing so, apart from enforcement of codes of behavior sanctioned by the local community and the student body. At the moment, there appears to be more general support for negative than for affirmative answers to these questions. But here again, only time, and the good judgment of well-informed educators, will tell.

[4]House Bill No. 5004

When Should the Tests Be Given?

In most of the early state testing programs, the tests were given in the spring at the end of the school year. This seemed a logical time to test for what had been learned. Then, partly in response to criticisms of end-of-year testing, some programs moved to fall or to midyear testing. Now, with renewed emphasis on assessing the results of instructional programs, there is some tendency to return to testing near the close of the school year.

The recent ETS survey showed that October was the month most frequently used for testing, followed by September, April, and May. Of course, some programs offer tests, usually of different kinds and for different purposes, more than once a year. It is interesting to note two programs in which tests are administered during the usual vacation months of July and August.

As suggested above, there is a relation between the purposes for testing and the time of year when the tests are given. Tests intended for guidance, for identification of individual problems and talents, or for placement and grouping are usually given in the fall. Those intended for evaluation of educational programs or instruction are most frequently given in the spring. Of course, any of these purposes can be served to some degree by tests given at any time of the year. But the timing of the testing does affect the extent to which a particular purpose is served well or poorly.

What Kinds of Tests Should Be Given?

The tests that actually are being used are mainly tests of basic skills in reading, mathematics, and language; of basic understandings in natural science and social science; of aptitudes, and of study skills. Many of these are standardized tests, available from commercial test publishers. But tests that were tailor-made, or especially revised, were used in a substantial minority of the states.

Standardized tests have a number of advantages. One, of course, is ready availability. The director of a state testing program can usually secure the tests and other needed materials such as answer sheets, directions for administration, manuals for score interpretation on relatively short notice. Once the structure of the program has been determined, the time required to get it into operation need not be long.

But this ready availability can sometimes be a disadvantage. Teachers with pupils to be tested may secure copies of the tests in advance and use them in "teaching" (coaching). Such practices seriously limit the validity of the tests as measures of real achievement and lead to grossly unfair comparisons between schools. Directors of state testing programs that make use of published tests must take care to see that no advance information on the test to be used is released.

A second valuable attribute of published standardized tests is their generally high quality. Usually the construction of a standardized test is directed by able, well trained, experienced test specialists. It is true, as Buros' mental measurement yearbooks attest, that when these test specialists look at tests constructed by other specialists, they can often point out shortcomings and suggest improvements. But most of the widely used published standardized tests are about as high in quality as the state of the art and the economic constraints of publishing allow.

A third advantage of standardized tests is that national norms frequently accompany them. These can sometimes supply useful information to supplement state and local norms, indicating how the educational achievements of pupils in a particular school or state compare with those of the nation as a whole.

A fourth characteristic of standardized tests is perhaps more commonly regarded as a disadvantage than as an advantage. The aspects of achievement covered in such tests are those most generally regarded as important. If they are not, the test is not likely to be widely used. The aspects covered may not correspond closely with those that are given greatest emphasis in a particular school or system. But if a substantial discrepancy exists, particularly in programs designed to develop the basic skills or to cultivate understanding in basic areas of knowledge, it may be the local program that is more to be questioned than the standard test. The latter probably has been developed on a broader basis of more expert judgment than has the local program. It is common for committees of expert teachers to have a hand in planning and developing a standardized test.

While a good case can be made for experimental innovations in methods of teaching, it is much harder to make a strong case for local uniqueness in goals of instruction in the common branches of learning. And even if a school does have somewhat unique ideas about what pupils should be learning, it probably is a good idea to find out how this program is affecting the achievement of what others regard as important.

Tailor-Made Tests

The advantages and drawbacks of tailor-made tests are roughly the reverse of those of standardized tests. Perhaps the most apparent and important advantage of the tailor-made test is that it can be designed to measure the educational outcomes judged by the program policy authorities to be most essential in their particular situation. In some cases, this can be an extremely important factor, but as was pointed out earlier, the worth of locally unique educational goals in the common branches of learning is open to some question.

Another advantage of the tailor-made test is that the security of the test can be more easily protected. Misuse of the test in coaching can be largely eliminated unless, of course, the same test is used repeatedly. When test security is protected, the validity of the test scores, and of intergroup comparisons, can be maintained.

A major problem in the use of tailor-made tests is finding good tailors. Few state departments of education, and, indeed, not all state universities, have staff members whose talent, training, and experience adequately qualify them to do a good job of test development.

Contracting with an agency that specializes in testing is probably the best solution to the problem of tailor-made test development. Often such agencies have files of tested items from which appropriate selections can be made. But this solution brings problems of its own.

There is the problem of defining, and of communicating to the testing agency, exactly what the contents and characteristics of the tests should be. There is the contractor's problem of meeting those specifications. And there is the difficult problem of fixing responsibility for the quality of the product when it is the product of joint efforts. The test user is likely not to be wholly satisfied with what the test producer gives him to use. This is not to say that cooperative test development is unworkable. But it is to say that the task is not as simple as it may appear at first glance.

It follows from this that the cost of a tailor-made test is likely to be higher than that of a published one. And, of course, only state and local norms can ordinarily be developed for a tailor-made test.

Thus, it appears that neither standardized nor tailor-made tests provide ideal answers to the question of what kind of tests should be given. But since no other answer seems to be available, one of the two, or a combination of both, may have to be chosen. It would be difficult to exaggerate the importance of test quality to the success of a state testing program. Other things—opposition of school officials to "external" testing, lack of funds, or inept management—may cause the program to languish and fail. But without appropriate tests of high quality it can have no hope of long run survival.

Other Problems of State Testing Programs

It is interesting to note that test quality was not mentioned as a major problem by directors of any of the state testing programs surveyed recently. The problem reported most frequently (by 11 states) was funding. As

mentioned earlier in this article, reduction of ESEA Title III funds may reduce or eliminate the testing programs in some states.

Funds for educational purposes are seldom supplied as generously as most educators would like. In the years immediately ahead, public generosity for educational purposes is likely to be somewhat less than it was during the quarter century from 1945 to 1970. But while lack of funds may be cited as the reason for the demise of some testing programs, the real reasons will probably not be financial stringency. Instead they will be that either:

1. The program did not yield clearly valid, easily interpretable, data on the effectiveness of educational efforts in the state, or that
2. The public interest in obtaining such data was not marshalled effectively enough to overcome the objections of educators to external evaluations.

If state testing programs disappear, it will not be primarily for lack of educational funds. It will be due more fundamentally to deficiencies in the wisdom and skill of the test specialists or to limitations in the vision and strength of educational leaders.

A second problem mentioned by several directors of state testing programs was "use of results," presumably inadequate or inappropriate use. No doubt, examples of these deficiencies could be cited. But the feeling that they constitute a major problem may be exaggerated.

The purpose of many state testing programs is simply to provide information: information that *can* be used as part of the basis for decision making in the legislature, the state department of education, the school board, or in the classroom; information that *will* be so used if it is relevant, reliable, and meaningfully reported. These uses are likely to be numerous and diverse. They are set in motion not by the production of the test data but by the recognition of an educational problem. One thing that a state testing program can do to promote effective use of results is to prepare a suggestive case book of appropriate uses that have been made of the results. But far more important and basic than this is to provide meaningful reports of relevant, reliable results to appropriate educational decision makers.

Future Prospects

What of the future of state testing programs? It is difficult at this point to predict with any degree of certainty whether they will flourish or languish. Clearly, there is a need for the kind of information on educational effectiveness that state testing programs can provide. That need is likely to continue and to grow. What is not clear is whether the leaders of state testing programs will be able to supply that information meaningfully and reliably and whether educational leaders will tolerate the immediate pain it sometimes brings as a necessary price to be paid for advancement of the enterprise to which they are committed. An end to attacks on testing appears unlikely in the forseeable future. But the growing public demand that assertions about the quality of education in a school be backed by solid evidence gives one grounds for hope that those who support the assessment of educational outcomes will prevail.

REFERENCES*

Segel, David. *State testing and evaluation programs.* Washington, D.C.: United States Office of Education, 1951.

State testing programs: a survey of functions, tests, materials and services. Princeton, N.J.: Educational Testing Service, 1968. ED 080 536

State educational assessment programs: 1973 revision. Princeton, N.J.: Educational Testing Service, 1973. ED 080 582

State educational assessment programs. Princeton, N.J.: Educational Testing Service, 1971. ED 056 102

State testing programs: 1973 revision. Princeton, N.J.: Educational Testing Service. ED 087 789

*Items followed by an ED number (for example ED 069 762) are available from the ERIC Document Reproduction Service (EDRS). Consult the most recent issue of *Resources in Education* for the address and ordering information.